Cross-lingual Induction of Selectional Preferences with Bilingual Vector Spaces

Yves Peirsman
QLVL, University of Leuven
Research Foundation – Flanders (FWO)
yves.peirsman@arts.kuleuven.be

Sebastian Padó
IMS, University of Stuttgart
pado@ims.uni-stuttgart.de

Abstract

We describe a cross-lingual method for the induction of selectional preferences for resource-poor languages, where no accurate monolingual models are available. The method uses bilingual vector spaces to "translate" foreign language predicate-argument structures into a resource-rich language like English. The only prerequisite for constructing the bilingual vector space is a large unparsed corpus in the resource-poor language, although the model can profit from (even noisy) syntactic knowledge. Our experiments show that the cross-lingual predictions correlate well with human ratings, clearly outperforming monolingual baseline models.

1 Introduction

Selectional preferences capture the empirical observation that not all words are equally good arguments to a given verb in a particular argument position (Wilks, 1975; Resnik, 1996). For instance, the subjects of the English verb *to shoot* are generally people, while the direct objects can be people or animals. This is reflected in speakers' intuitions. Table 1 shows that the combination *the hunter shot the deer* is judged more plausible than *the deer shot the hunter*. Selectional preferences do not only play an important role in human sentence processing (McRae et al., 1998), but are also helpful for NLP tasks like word sense disambiguation (McCarthy and Carroll, 2003) and semantic role labeling (Gildea and Jurafsky, 2002).

Computational models of selectional preferences predict such *plausibilities* for triples of a predicate p, an argument position a, and a head word h, such as (*shoot*,*object*,*hunter*). All recent models take a two-step approach: (1), they extract all triples (p, a, h) from a large corpus; (2), they apply some type of generalization to make predictions for unseen items. Clearly, the accuracy of these models relies crucially on the quality and coverage of the extracted triples, and thus on the syntactic analysis of the corpus. Unfortunately, corpora that are both large enough and have a very good syntactic analysis are only available for a handful of Western and Asian languages, which leaves all other languages without reliable selectional preference models.

Predicate	Relation	Noun	Plausibility
shoot	subject	hunter	6.9
shoot	object	hunter	2.8
shoot	subject	deer	1.0
shoot	object	deer	6.4

Table 1: Predicate-relation-noun triples with human plausibility judgments on a 7-point scale (McRae et al., 1998)

In this paper, we propose a cross-lingual knowledge transfer approach to this problem: We automatically *translate* triples (p, a, h) from resource-poor languages into English, where large and high-quality parsed corpora are available and we can compute a reliable plausibility estimate. The translations are extracted from a *bilingual semantic space*, which can be constructed via bootstrapping from large unparsed corpora in the two languages, without the need for parallel corpora or bilingual lexical resources.

Structure of the paper. Section 2 reviews models for selectional preferences. In Section 3, we describe our approach. Section 4 introduces our experimental setup, and Sections 5 and 6 present and discuss our experiments. Section 7 wraps up.

2 Selectional Preferences

The first broad-coverage model of selectional preferences was developed by Resnik (1996). To estimate the plausibility of a triple (p, a, h), Resnik first extracted all head words seen with predicate p in position a, $Seen_a(p)$, from a corpus. He then used the WordNet hierarchy to generalize over the head words and to create predictions for unseen ones. A number of studies has followed the same approach, exploring different ways of using the structure of WordNet (Abe and Li, 1996; Clark and Weir, 2002). While these approaches show good results, they can only make predictions for argument heads that are covered by WordNet. This is already a problem for English, and much more so in other languages, where comparable resources are often much smaller or entirely absent.

A promising alternative approach is to derive the generalizations from distributional information (Prescher et al., 2000; Padó et al., 2007; Bergsma et al., 2008). For example, the Padó et al. (2007) model computes vector space representations for all head words h and defines the plausibility of the triple (p, a, h) as a weighted mean of the vector space similarities between h and all h' in $Seen_a(p)$:

$$Pl(p, a, h) = \sum_{h' \in Seen_a(p)} \frac{w(h') \cdot sim(h, h')}{\sum_{h'} w(h')} \quad (1)$$

where $w(h')$ is a weight, typically frequency.

In this model, the generalization is provided by distributional similarity, which can be computed from a large corpus, without the need for additional lexical resources. Padó et al. found it to outperform Resnik's approach in an evaluation against human plausibility judgments. However, note that competitive results are only obtained by representing the head words in "syntactic" vector spaces whose dimensions consist of context words with their syntactic relation to the target rather than just context words. This is not surprising: Presumably, *hunter* and *deer* share a domain and are likely to have similar word-based context distributions, even though they differ with regard to their plausibility for particular predicate-argument positions. Only when the vector space can capture their different *syntactic* co-occurrence patterns can the model predict different plausibilities.

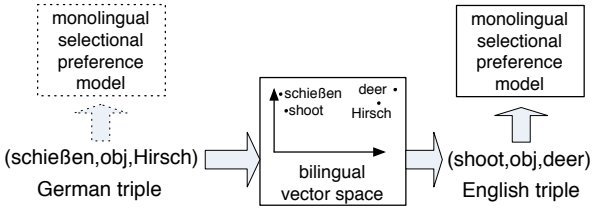

Figure 1: Predicting selectional preferences for a source language (e.g. German) by translating into a target language (e.g. English) with a bilingual vector space.

3 Cross-lingual selectional preferences

In order to compute reliable selectional preference representations, distributional models need to see at least some head words for each (p, a) combination. Manually annotated treebank corpora, which are becoming available for an increasing number of languages, are too small for this task. We therefore explore the idea of predicting the selectional preferences for such languages by taking advantage of large corpora with high-quality syntactic analyses in resource-rich languages like English. This idea falls into the general approach of *cross-lingual knowledge transfer* (see e.g. Hwa et al., 2005). The application to selectional preferences was suggested by Agirre et al. (2003), who demonstrated its feasibility by manual translation between Basque and English. We extend their experiments to an automatic model that predicts plausibility judgments in a resource-poor language (*source language*) by exploiting a model in a resource-rich language (*target language*).

Figure 1 sketches our method. We assume that there is not enough high-quality data to build a monolingual selectional preference model for the source language (shown by dotted lines). However, we can use a *bilingual vector space*, that is, a semantic space in which words of both the source and the target language are represented, to *translate* each source language word s into the target language by identifying its nearest (most similar) target word $tr(s)$:

$$tr(s) = \text{argmax}_t \, sim(s, t) \quad (2)$$

Now we can use a target language selectional preference model to obtain plausibilities for source triples:

$$Pl^s(p, a, h) = Pl^t(tr(p), a, tr(h)) \quad (3)$$

where the superscript indicates the language.

Eq. (3) gives rise to three questions: (1), How can we construct the bilingual space to model *tr*? (2), Is translating actually the appropriate way of transferring selectional preferences? (3), Is it reasonable to retain the source language argument positions like *subject* or *object*? The following subsections discuss (1) and (2); we will address (3) in Sections 5 and 6.

3.1 Bilingual Vector Spaces

Bilingual vector spaces are vector spaces in which words from two languages are represented (cf. Fig. 2). The dimensions of this space are labeled with bilingual context word pairs (like *secretly/heimlich* and *rifle/Gewehr* for German–English) that are mutual translations. By treating such context word pairs as single dimensions, the vector space can represent target words from both languages, counting the target words' co-occurrences with the context words from the respective language. In other words, a source-target word pair (s, t) will be assigned similar vectors in the semantic space if the context words of s are translations of the context words of t. Cross-lingual semantic similarity between words can be measured using standard vector space similarity (Lee, 1999).

Importantly, bilingual vector spaces can be built on the basis of co-occurrences drawn from two unrelated corpora for the source and target languages. Their construction does not require resources such as parallel corpora or bilingual translation lexicons, which might not be available for resource-poor source languages. Where parallel corpora exist, they often cover specific domains (e.g., politics), while many bilingual lexicons are prone to ambiguity problems.

The main challenge in constructing bilingual vector spaces is determining the set of dimensions, i.e., bilingual word pairs, using as little knowledge as possible. Most often, such pairs are extracted from small bilingual lexicons (Fung and McKeown, 1997; Rapp, 1999; Chiao and Zweigenbaum, 2002). As mentioned above, such resources might not be available. We thus follow an alternative approach by using frequent *cognates*, words that are shared between the two languages (Markó et al., 2005). Cognates can be extracted by simple string matching between the corpora, and mostly share their meaning (Koehn and Knight, 2002). However, they account for (at most) a small percentage of all interesting translation pairs.

To extend the set of dimensions available for the bilingual space, we use these cognates merely as a starting point for a *bootstrapping process*: We build a bilingual vector space with the initial word pairs as dimensions, and identify nearest neighbors between the two languages in the space. These are added as dimensions of the bilingual space, and the process is repeated. Since the focus is on identifying reliable source-target word pairs rather than complete coverage as in Eq. (2), we adopt a *symmetrical* definition of translation that pairs up only mutual nearest neighbors, and allows words to remain untranslated:[1]

$$tr_{sym}(s) = t \text{ iff } tr(s) = t \text{ and } tr(t) = s \qquad (4)$$

From the second iteration onward, this process introduces dimensions that are not identical graphemes, such as *Kind–child* and *Geschwindigkeit–speed*, and is iterated until convergence. Since each word of either language can only participate in at most one dimension, dimensions acquired in later steps can correct wrong pairs from previous steps, like the "false friend" German *Kind* 'child' – English *kind*, which is part of the initial set of cognates.

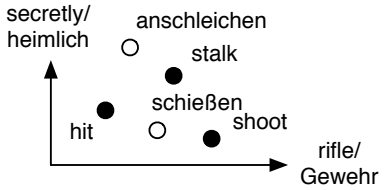

Figure 2: Sketch of a bilingual vector space for English (solid dots) and German (empty circles).

3.2 Translation and Selectional Preferences

As Figure 1 shows, the easiest way of exploiting a bilingual semantic space is to identify for each source word the target language word with the highest semantic similarity. For example, in Figure 2, the best translation of German *schießen* is its English nearest neighbor, *shoot*. However, it is risky to rely on the single nearest neighbor – it might simply be wrong. Even if it is correct, *data sparsity* is an issue: The translations may be infrequent in the target language, or the two translations of p and h may form unlikely collocates for target language-internal reasons (like

[1] To avoid unreliable vectors, we also adopt only the 50% most frequent of the tr_{sym} pairs. Frequency is defined as the geometric mean of the two words' monolingual frequencies.

difference in register) that do not reflect plausibility. A third issue are monolingual semantic phenomena like *polysemy* and *idioms*: The implausible German triple *(schießen,obj,Brise)* will be judged as very plausible due to the English idiom *to shoot the breeze*.

A look at the broader neighborhood of *schießen* suggests that its second and third-best English neighbors, *hit*, and *stalk*, can be used to *smooth* plausibility estimates for *schießen*. Instead of translating source language words by their single nearest neighbor, we will take its k nearest neighbors into account. This is defensible also from a more fundamental point of view, which suggests that the cross-lingual transfer of selectional preferences does *not* require literal translation in order to work. First, ontological models like Resnik's assume that *synonymous* words behave similarly with respect to selectional preferences. Second, recent work by Chambers and Jurafsky (2009) has induced "narrative chains", i.e., likely sequences of events, by their use of similar head words. Thus, we expect that all k nearest neighbors of a source predicate s are *informative* for the selectional preferences of s (like *schießen*) as long as they are either synonyms of its literal translation (*shoot/hit*) or come from the same narrative chain (*stalk/kill/...*).

It is also clear that smoothing does not always equate better predictions. Closeness in a word-based vector space can also just reflect semantic association. For example, Spanish *tenista* 'tennis player' is highly associated with English *tennis*, but is a bad translation in terms of selectional preferences. We assume that this problem is more acute for nouns than for verbs: The context of verbs is dominated by their arguments, which is not true for nouns. Consequently, close nouns in vector space can differ widely in ontological type, while close verbs generally have one or more similar argument slots. In our model, we will thus consider several verb translations, but just the best head word translation. For details, see Section 5.

4 Experimental Setup

Our evaluation uses English as the target language and two source languages: German (as a very close neighbor of English) and Spanish (as a more distant one). Neither of these languages are really resource-poor, but they allow us to compare our cross-lingual model against monolingual models, to emulate different levels of "resource poorness" and to examine the model's learning curve.

Plausibility Data. For German, we used the plausibility judgments collected by Brockmann (2002). The dataset contains human judgments for ninety triples sampled from the manually annotated 1 million word TiGer corpus (Brants et al., 2002): ten verbs with three argument positions (subject [SUBJ], direct object [DOBJ], and oblique (prepositional) object [POBJ]) combined with three head words. Models are evaluated against such datasets by correlating predicted plausibilities with the (not normally distributed) human judgments using Spearman's ρ, a non-parametric rank-order correlation coefficient.

We constructed a similar 90-triple data set for Spanish by sampling triples from two Spanish corpora (see below) using Brockmann's (2002) criteria. Human judgments for the triples were collected through the Amazon Mechanical Turk (AMT) crowdsourcing platform (Snow et al., 2008). We asked native speakers of Spanish to rate the plausibility of a simple sentence with the relevant verb-argument combination on a five-point Likert scale, obtaining between 12 and 17 judgments for each triple. For each datapoint, we removed the single lowest and highest judgments and computed the mean. We assessed the reliability of our data by replicating Brockmann's experiment for German with our AMT setup. With a Spearman ρ of almost .90, our own judgments correlate very well with Brockmann's original data.

Monolingual Prior Work and Baselines. For German, Brockmann and Lapata (2003) evaluated ontology-based models trained on TiGer triples and the GermaNet ontology. The results in Table 2 show that while both models are able to predict the data significantly, neither of the models can predict all of the data. We attribute this to the small size of TiGer.[2]

To gauge the limits of monolingual knowledge-lean approaches, we constructed two monolingual distributional models for German and Spanish according to the Padó et al. (2007) model (Eq. (1)). Recall that this model performs generalization in a syntax-based vector space model. We computed vector spaces from dependency-parsed corpora for the

[2]For each of the three argument positions and "all", Brockmann and Lapata report the results for the best parametrization of the models, which explains the apparently inconsistent results.

	Resnik	Clark & Weir
SUBJ	.408*	.268
DOBJ	.430*	.611***
POBJ	.330	.597***
all	.374***	.232*

Table 2: Monolingual baselines 1. Spearman correlations for ontology-based models in German as reported by Brockmann and Lapata (2003). *: $p < .05$; ***: $p < .001$

Lang.	German		Spanish			
Corpus	Schulte's HGC		AnCora		Encarta	
	ρ	Cov.	ρ	Cov.	ρ	Cov.
SUBJ	.34†	90%	.44*	80%	.14	100%
DOBJ	.51**	97%	.29	83%	-.05	100%
POBJ	.41*	93%	-.03	100%	—	—[3]
all	.33**	93%	.16	88%	.11	67%

Table 3: Monolingual baselines 2. Spearman correlation and coverage for distributional models. † : $p < .1$; *: $p < .05$; **: $p < .01$.

two languages, using the 2,000 most frequent lemma-dependency relation pairs as dimensions and adopting the popular pointwise mutual information metric as co-occurrence statistic. For German, we used Schulte im Walde's verb frame resource (Schulte im Walde et al., 2001), which contains the frequency of triples calculated from probabilistic parses of 30M words from the Huge German Corpus (HGC) of newswire. For Spanish, we consulted two syntactically analyzed corpora: the AnCora (Taulé et al., 2008) and the Encarta corpus (Calvo et al., 2005). At 0.5M words, the AnCora corpus is small, but manually annotated, whereas the larger, automatically parsed Encarta corpus amounts to over 18M tokens.

Table 3 shows the results for the distributional monolingual models. For German, we get significant correlations for DOBJ and POBJ, an almost significant correlation for SUBJs, and high significance for the complete dataset ($p < 0.01$). These figures rival the performance of the ontological models (cf. Table 2), without using ontological information. For Spanish, the only significant correlation with human judgments is obtained for subjects, the most frequent argument position, with the clean AnCora data. AnCora is presumably too sparse for the other argument positions. The large Encarta corpus, in turn, is very noisy, supporting our concerns from Section 2.

[3]Since the Encarta data consists of individual dependency relations rather than trees, we could not model the POBJ data.

	n	noun	adj	verb	all
German	7340	.61	.57	.43	.56
Spanish	4143	.62	.67	.41	.58

Table 4: First-translation accuracy for German-English and Spanish-English translation (n: size of gold standard).

Cross-lingual Selectional Preferences. Our architecture for the cross-lingual prediction of selectional preferences shown in Figure 1 consists of two components, namely the bilingual vector space and a selectional preference model in the target language.

As our English selectional preference model, we again use the Padó et al. (2007) model, trained on a version of the BNC parsed with MINIPAR (Lin, 1993). The parameters of the syntactic vector space were the same as for the monolingual baseline models. The bilingual vector spaces were constructed from three large, unparsed, comparable monolingual corpora. For German, we used the HGC described above. For Spanish, we obtained a corpus with around 100M words, consisting of 2.5 years of crawled text from two major Spanish newspapers. For English, we used the BNC.

We first constructed initial sets of bilingual labels. For German–English, we identified 1064 graphemically identical word pairs that occurred more than 4 times per million words. Due to the larger lexical distance between Spanish and English, there are fewer graphemically identical tokens for this language pair. We therefore applied a Porter stemmer and found 2104 identical stems, at a higher risk of "false friends". We then applied the bootstrapping cycle from Section 3.1. The set of dimensions converged after around five iterations.

We evaluated the (asymmetric) nearest neighbor pairs from the final spaces, $(s, tr(s))$, against two online dictionaries.[4] Table 4 shows that 55% to 60% of the pairs are listed in the dictionaries, with parallel tendencies for both language pairs. The bilingual space performs fairly well for nouns and adjectives, but badly for verbs, which is a well-known weakness of distributional models (Peirsman et al., 2008).

Even taking into account the incompleteness of dictionaries, this looks like a negative result: more

[4]DE-EN: `www.dict.cc`; ES-EN: `www.freelang.net`. Pairs $(s, tr(s))$ were only evaluated if the dictionary listed s.

than half of all verb translations are incorrect. However, following up on our intuitions from Section 3.2, we performed an analysis of the "incorrect" translations. It revealed that many of the errors in Table 4 are informative, semantically related words. Nearest neighbor target language verbs in particular tend to represent the same event type and take the same kinds of arguments as the source verb. Examples are German *gefährden* 'threaten' – English *affect*, and German *Neugier* 'curiosity' – English *enthusiasm*. We concluded that literal translation quality is a misleading figure of merit for our task.

Experimental rationale. Section 3 introduced one major design decision of our model: the question of how to treat the *argument position*, which cannot be translated by the bilingual vector space, in the cross-lingual transfer. We present two experiments that investigate the model's behavior in the absence and presence of knowledge about argument positions. Experiment 1 uses no syntactic knowledge about the source language whatsoever. In this situation, the best we can do is to assume that source language argument positions like SUBJ will correspond to the same argument position in the target language. Experiment 2 attempts to identify, for each source language argument position, the "best fit" position in the target language. This results in better plausibility estimates, but also means that we need at least some syntactic information about the source language. In both experiments, we vary the number of translations we consider for each verb.

5 Exp. 1: Induction without syntactic knowledge in the source language

This experiment assumes that argument positions simply carry over between languages. While this assumption clearly simplifies linguistic reality, it has the advantage of not needing any syntactic information about the source language. We thus model German and Spanish SUBJ relations by English SUBJ relations and DOBJs by DOBJs. In the case of (lexicalized) POBJs, where we cannot assume identity, we compute plausibility scores for *all* English POBJs that account for at least 10% of the predicate's argument tokens, and select the PP with the highest plausibility estimate. The k best "translations" of the predicate p, $tr_k(p)$, are turned into a single prediction using maximization, yielding the final model:

$$Pl^s_{\text{nosyn}}(p, a, h) = \max_{p_t \in tr_k(p)} Pl^t(p_t, a, tr(h)) \quad (5)$$

Note that this model does not use any source language information, except the bilingual vector space.

The results of Experiment 1 are given in Table 5 (coverage always 100%). For German, all predictions correlate significantly with human ratings, and most even at $p < 0.01$, despite our naive assumption about the cross-lingual argument position identity. The results exceed both monolingual model types (ontological, Tab. 2, and distributional, Tab. 3), notably without the use of syntactic data. In particular, the results for the POBJs, notoriously difficult to model monolingually, are higher than for SUBJs or DOBJs. We attribute this to the cross-lingual generalization which takes all prepositional arguments into account.

The Spanish dataset is harder to model overall. We obtain significantly high correlations for SUBJ, but non-significant results for DOBJ and POBJ. This corresponds well to the patterns for the monolingual AnCora corpus (Table 3). However, we outperform AnCora on the complete dataset, where it did not achieve significance, while the cross-lingual model does at $p < 0.01$ — again, even without the use of syntactic analyses. We attribute the overall lower results compared to German to systematic syntactic differences between English and Spanish. For example, animate direct objects in Spanish are realized as POBJs headed by the preposition *a*. Estimating the plausibility of such objects by looking at English POBJs is unlikely to yield good results. The use of a larger number of verb translations yields a clear increase in correlation for the German data, but inconclusive results for Spanish.

6 Exp. 2: Induction with syntactic knowledge in the source language

As discussed in Section 3.2, verbs that are semantically similar in the bilingual vector space may very well realize their (semantic) argument positions differently in the surface syntax. For example, German *teilnehmen* is correctly translated to English *attend*, but the crucial event argument is realized differently, namely as a POBJ headed by *an* in German and as a DOBJ in English. To address this problem, we

DE	1-best	2-best	3-best	4-best	5-best
SUBJ	.44*	.47**	.45*	.47**	**.54****
DOBJ	.39*	.39*	.52**	.54**	**.55****
POBJ	.58**	.61**	.61**	.61**	**.62****
all	.35**	.37**	.37**	.38**	**.40****

ES	1-best	2-best	3-best	4-best	5-best
SUBJ	.58**	**.64****	**.64****	.58**	.58**
DOBJ	.13	**.16**	.11	.07	.07
POBJ	**.13**	**.13**	.09	.14	.14
all	.34**	**.36****	.34**	.32**	.32**

Table 5: Exp.1: Spearman correlation between syntaxless cross-lingual model and human judgments for k best verb translations. Best k for each argument position marked in boldface. Coverage of all models: 100%.

DE	1-best	2-best	3-best	4-best	5-best
SUBJ	.55**	**.59****	.49**	.52**	.54**
DOBJ	.52**	.52**	.66**	.66**	**.68****
POBJ	.61**	.68**	**.70****	.69**	**.70****
all	.41**	.44**	.44*	.46**	**.48****

ES-A	1-best	2-best	3-best	4-best	5-best
SUBJ	**.52****	.47*	.42*	.41*	.42*
DOBJ	.52*^{c}	**.64**^{c}**	.54*^{c}	.42*^{c}	.42*^{c}
POBJ	**.32**†**	.18	.13	.13	.24
all	**.47****	.41**	.36**	.33**	.37**

ES-E	1-best	2-best	3-best	4-best	5-best
SUBJ	.40*	**.42***	.39*	.39*	.41*
DOBJ	**.21**	.02	.06	.13	.20

Table 6: Exp.2: Spearman correlation between syntax-aware cross-lingual model and human judgments for k best verb translations. ES-A: AnCora corpus, ES-E: Encarta corpus. Best k for each argument position in boldface. Coverage of all models: 100%, except ^{c}: 60%.

learn a mapping function m that identifies the argument position a_t of a target language predicate p_t that corresponds best to an argument position a of a predicate p in the source language. Our simple model is in the same spirit as the cross-lingual plausibility model itself: It returns the argument position a_t of p_t for which the seen head words of (p, a) are most plausible when translated into the target language:[5]

$$m(p, a, p_t) = \operatorname*{argmax}_{a_t} \sum_{h \in Seen_a(p)} Pl^t(p_t, a_t, tr(h))$$

Parallel to Eq. (5), the cross-lingual model is now:

$$Pl^s_{\text{syn}}(p, a, h) = \max_{p_t \in tr_k(p)} Pl^t(p_t, m(p, a, p_t), tr(h)) \quad (6)$$

This model can recover English argument positions that correspond better to the original ones than the identity mapping. For example, on our data, it discovers the mapping for *teilnehmen an/attend* discussed above. A second example concerns the incorrect, but informative translation of *stagnieren* 'stagnate' as *boost*. Here the model recognizes that the SUBJ of *stagnieren* (the stagnating entity) corresponds to the DOBJ of *boost*.

Establishing m requires syntactic information in the source language, in order to obtain the set of seen head words $Seen_{a_s}(p_s)$. For this reason, Exp. 2 uses the parsed subset of the HGC (German), and the AnCora and Encarta corpora (Spanish). The results are shown in Table 6. We generally improve over Exp. 1. For German, every single model now correlates highly significantly with human judgments ($p < 0.01$), and the correlation for the complete dataset increases from .40 to .48. For Spanish, we see very good results for the AnCora corpus. Compared to Exp. 1, we see a slight degradation for the SUBJs; however, the correlations remain significant for all values of k. Conversely, all predictions for DOBJs are now significant,[6] and the POBJs have improved at least numerically, which validates our analysis of the problems in Exp. 1. The best correlation for the complete dataset improves from .36 to .47. The results for the Encarta corpus disappoint, though. SUBJs are significant, but worse than for AnCora, and the DOBJs remain non-significant throughout. With regard to increasing the number of verb translations, Exp. 2 shows an almost universal benefit for German, but still mixed results for Spanish, which may indicate that verb translations for Spanish are still "looser" than the German ones.

In fact, most remaining poor judgments are the result of problematic translations, which stem from three main sources. The first one is sparse data. Infrequent German and Spanish words often receive unreliable vector representations. Some examples are the

[5]To alleviate sparse data, we ignore argument positions of English verbs that represent less than 10% of its argument tokens.

[6]Note, however, that AnCora has an imperfect coverage for DOBJs (60%). This is because our Spanish dataset contains verbs sampled from Encarta that do not occur in AnCora.

German *Tau* ('dew', frequency of 180 in the HGC), translated as *alley*, and *Reifeprüfung* (German SAT, frequency 120), translated as *affiliation*. Both of these may also be due to the difference in genre between the HGC and the BNC. A second problem is formed by nearest neighbors that are ontologically dissimilar, as in the *tenista* 'tennis player'/*tennis* example from above. A final issue relates to limitations of the Padó et al. (2007) model, whose architecture is susceptible to polysemy-related problems. For instance, the Spanish combination (*excavar*, *obj*, *terreno*) was judged by speakers as very plausible, but its English equivalent (*excavate*, *obj*, *land*) is assigned a very low score by the model. This might be due to the fact that in the BNC, *land* occurs often in its political meaning, and forms an outlier among the head words for (*excavate*,*obj*).

How much syntactic information is necessary? The syntax-aware model requires syntactic information about the source language, which seems to run counter to our original motivation of developing methods for resource-poor languages. To address this point, we analyzed the behavior of the syntax-aware model for small syntactically analyzed corpora that contained only at most m occurrences for each predicate. We obtained the m occurrences by sampling from the syntactically analyzed part of the HGC; if fewer than m occurrences were present in the corpus, we simply used these. Figure 3 shows the training curve with 1 verb translation, averaged over n rounds ($n = 10$ for 5 arguments, $n = 5$ for 10 arguments, $n = 4$ for 20, 50 and 100 arguments). The general picture is clear: most of the benefit of the syntactic data is drawn form the first five occurrences for each argument position. This shows that a small amount of targeted syntactic annotation can improve the cross-lingual model substantially.

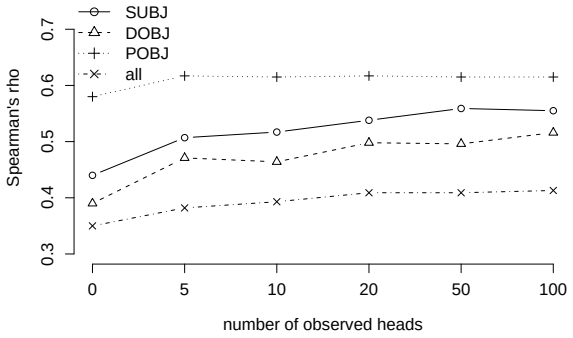

Figure 3: Training curve for the bilingual German–English model as a function of the number of observed head words per argument position in the source language.

7 Conclusions

In this article, we have presented a first unsupervised cross-lingual model of selectional preferences. Our model proceeds by automatically *translating* (predicate, argument position, head word) triples for resource-poor source languages into a resource-rich target language, where accurate selectional preference models are available. The translation is based on a bilingual vector space, which can be bootstrapped from large unparsed corpora in the two languages.

Our results indicate that bilingual methods can go a long way towards the modeling of selectional preferences in resource-poor languages, where bilingual lexicons, parallel corpora, or ontologies might not be available. Our experiments have looked at German and Spanish, where the cross-lingual models rival and even exceed monolingual methods that typically have to rely on small, clean "treebank"-style corpora or large, very noisy, automatically parsed corpora. We have also demonstrated that noisy syntactic data from the source language can be integrated in our model, where it helps improve the cross-lingual handling of argument positions. The linguistic distance between the languages can impact (1) the ability to find accurate translations and (2) the degree of syntactic overlap; nevertheless, as Agirre et al. (2003) show, the transfer is possible even for unrelated languages.

In this paper, we have instantiated the selectional preference model in the target language (English) with the distributional model by Padó et al. (2007). However, our approach is modular and can be combined with any other selectional preference model. We see two main avenues for future work: (1), The construction of properly bilingual models where source language information can also help to further improve the *target* language model (Diab and Resnik, 2002); (2), The extension of our cross-lingual mapping for the argument position to mappings that hold across multiple predicates as well as argument-dependent mappings like the Spanish direct objects, whose realization depends on their animacy.

References

Naoki Abe and Hang Li. 1996. Learning word association norms using tree cut pair models. In *Proc. ICML*, pages 3–11, Bari, Italy.

Eneko Agirre, Izaskun Aldezabal, and Eli Pociello. 2003. A pilot study of English selectional preferences and their cross-lingual compatibility with Basque. In *Proc. TSD*, pages 12–19, Brno, Czech Republic.

Shane Bergsma, Dekang Lin, and Randy Goebel. 2008. Discriminative learning of selectional preference from unlabeled text. In *Proc. EMNLP*, pages 59–68, Honolulu, HI.

Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. The TIGER treebank. In *Proc. Workshop on Treebanks and Linguistic Theories*, Sozopol, Bulgaria.

Carsten Brockmann and Mirella Lapata. 2003. Evaluating and combining approaches to selectional preference acquisition. In *Proc. EACL*, pages 27–34, Budapest, Hungary.

Carsten Brockmann. 2002. Evaluating and combining approaches to selectional preference acquisition. Master's thesis, Universität des Saarlandes, Saarbrücken.

Hiram Calvo, Alexander Gelbukh, and Adam Kilgarriff. 2005. Distributional thesaurus vs. wordnet: A comparison of backoff techniques for unsupervised PP attachment. In *Proc. CICLing*, pages 177–188, Mexico City, Mexico.

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Proc. ACL*, pages 602–610, Singapore.

Yun-Chuang Chiao and Pierre Zweigenbaum. 2002. Looking for candidate translational equivalents in specialized, comparable corpora. In *Proc. COLING*, pages 1–5, Taipei, Taiwan.

Stephen Clark and David Weir. 2002. Class-based probability estimation using a semantic hierarchy. *Computational Linguistics*, 28(2):187–206.

Mona Diab and Philip Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. In *Proc. ACL*, pages 255–262, Philadelphia, PA.

Pascale Fung and Kathleen McKeown. 1997. Finding terminology translations from non-parallel corpora. In *Proc. 3rd Annual Workshop on Very Large Corpora*, pages 192–202, Hong Kong.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Rebecca Hwa, Philipp Resnik, Amy Weinberg, Clara Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, 11(3):311–325.

Philipp Koehn and Kevin Knight. 2002. Learning a translation lexicon from monolingual corpora. In *Proc. ACL-02 Workshop on Unsupervised Lexical Acquisition*, pages 9–16, Philadelphia, PA.

Lillian Lee. 1999. Measures of distributional similarity. In *Proc. ACL*, pages 25–32, College Park, MD.

Dekang Lin. 1993. Principle-based parsing without overgeneration. In *Proc. ACL*, pages 112–120.

Kornél Markó, Stefan Schulz, Olena Medelyan, and Udo Hahn. 2005. Bootstrapping dictionaries for cross-language information retrieval. In *Proc. SIGIR*, pages 528–535, Seattle, WA.

Diana McCarthy and John Carroll. 2003. Disambiguating nouns, verbs and adjectives using automatically acquired selectional preferences. *Computational Linguistics*, 29(4):639–654.

Ken McRae, Michael Spivey-Knowlton, and Michael Tanenhaus. 1998. Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, 38:283–312.

Sebastian Padó, Ulrike Padó, and Katrin Erk. 2007. Flexible, corpus-based modelling of human plausibility judgements. In *Proc. EMNLP-CoNLL*, pages 400–409, Prague, Czech Republic.

Yves Peirsman, Kris Heylen, and Dirk Geeraerts. 2008. Size matters. Tight and loose context definitions in English word space models. In *Proc. ESSLLI Workshop on Lexical Semantics*, pages 9–16, Hamburg, Germany.

Detlef Prescher, Stefan Riezler, and Mats Rooth. 2000. Using a probabilistic class-based lexicon for lexical ambiguity resolution. In *Proc. COLING*, pages 649–655, Saarbrücken, Germany.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Proc. ACL*, pages 519–526, College Park, MD.

Philip Resnik. 1996. Selectional constraints: An information-theoretic model and its computational realization. *Cognition*, 61:127–159.

Sabine Schulte im Walde, Helmut Schmid, Mats Rooth, Stefan Riezler, and Detlef Prescher. 2001. Statistical Grammar Models and Lexicon Acquisition. In *Linguistic Form and its Computation*, pages 389–440. CSLI Publications, Stanford, CA.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast – but is it good? Evaluating non-expert annotations for natural language tasks. In *Proc. EMNLP*, pages 254–263, Honolulu, HI.

Mariona Taulé, M. Antònia Martí, and Marta Recasens. 2008. Ancora: Multilevel annotated corpora for Catalan and Spanish. In *Proc. LREC*, Marrakech, Morocco.

Yorick Wilks. 1975. Preference semantics. In E. Keenan, editor, *Formal Semantics of Natural Language*. Cambridge University Press.